Preference learning with evolutionary Multivariate Adaptive Regression Spline model

Abou-Zleikha, Mohamed; Shaker, Noor; Christensen, Mads Græsbøll

Published in:
IEEE Congress on Evolutionary Computation (CEC)

DOI (link to publication from Publisher):
10.1109/CEC.2015.7257154

Publication date:
2015

Document Version
Accepted author manuscript, peer reviewed version

Preference Learning with Evolutionary-Based Multivariate Adaptive Regression Spline Model

Mohamed Abou-Zleikha
Audio Analysis Lab, ad:mt,
Aalborg University
Aalborg, Denmark
Email: moa@create.aau.dk

Noor Shaker
Center for Culture and Games
IT University of Copenhagen
Copenhagen, Denmark
Email: nosh@itu.dk

Mads Græsbøll Christensen
Audio Analysis Lab, ad:mt,
Aalborg University
Aalborg, Denmark
Email: mgc@create.aau.dk

***Abstract*—Collecting users' feedback through pairwise preferences is recommended over other ranking approaches as this method is more appealing for human decision making. Learning models from pairwise preference data is however an NP-hard problem. Therefore, constructing models that can effectively learn such data is a challenging task. Models are usually constructed with accuracy being the most important factor. Another vitally important aspect that is usually given less attention is expressiveness, i.e. how easy it is to explain the relationship between the model input and output. Most machine learning techniques are on either side of the performance-expressiveness spectrum especially when it comes to learning complex non-linear functions. This paper introduces a novel approach for pairwise preference learning through combining an evolutionary method with Multivariate Adaptive Regression Spline (MARS). MARS has the advantage of being a powerful method for function approximation as well as being relatively easy to interpret. This work evolve MARS models based on their efficiency in learning pairwise data. The method is tested on two datasets that collectively provide pairwise preference data of five cognitive states expressed by users. The method is analysed in terms of the performance, expressiveness and complexity and showed promising results in all aspects.**

I. Introduction

Preference Learning (PL) is a subfield in machine learning that has attracted increasing attention in Artificial Intelligence research in recent years. PL refers to the problem of learning a predictive preference models from observations [1]. Observations are usually empirical data collected from users's feedback as preferences or abstract utilities. Consequently, modelling preferences can then be tackled using utility functions or pairwise preference learning. Utility functions assign a numerical or ordinal utility to each instance and, thereafter, the problem of learning such functions becomes one of regression learning or ordered classification [1]. Pairwise preference learning, on the other hand, relies on the principle of comparing pairs of alternatives making the construction of an efficient predictor a difficult task. This is mainly because the data is not necessarily transitive and, therefore, ranking can not be usually defined in a unique way [1]. Hence, learning a global ranking function from pairwise preferences is an NP-hard problem [2] since one needs to find a ranking that is maximally consistent with the given binary preferences.

Pairwise preference learning is recommended over traditional ranking methods when collecting data from human users. This is mainly because the comparative approach is more intuitively appealing for human decision making [1]. Several studies could be found in the literature on learning models of pairwise preferences [3], [4], [5]. In this paper, we are interested in learning such models for cognitive modelling, a domain where several approaches have been investigated with varying results [5], [6], [7]. We propose the use of Multivariate Adaptive Regression Spline (MARS) [8] for preference learning. Our choice is motivated by its several desired advantages: MARS has shown promising results when solving regression problems that are competitive with neural networks and support vector regression [9]; this method has the advantage of being easy to understand and interpret compared to the other approaches [10]; MARS also has superiority over other partitioning approaches such as decision stump [11] when dealing with numerical values; the method can effectively handle non-linear data; and finally, one of its important features is that it performs an automatic feature selection. These properties make MARS well suited for our problem which entails the construction of accurate and understandable predictors of pairwise preferences from complex data.

As we are dealing with pairwise preference data where output values are defined for pairs and not for individual instances, we propose the use of artificial evolution to train our MARS models. We are not aware of any previous attempts on training MARS models on pairwise preference data, and we know about only one study that uses a Genetic Algorithm-based (GA) approach to train MARS models for a regression task [12]. Our approach, however, offers a number of improvements. In particular, while the previous study [12] evolves the features used for modelling and the number of the basis functions, our approach optimises these two factors as well as the other basis functions parameters which greatly affect the modelling performance. Furthermore, since we are dealing with data of pairwise nature, a new definition of the error function for the evolution process that takes into account the nature of the data is proposed. Finally, while the previous study uses a standard GA method, we investigate the use of Grammatical Evolution (GE) for evolving MARS preference models. GE offers a number of advantages over traditional GA methods: it permits an easy way of describing the individuals in the population (through the use of design grammar) making the problem easy to define and allowing an easy way to understand and interpret the results.

To analyse the performance of the proposed method and to compare it with the state-of-the-art approaches, we use

two different datasets. The datasets contains instances of pairwise preference data collected from users' feedback in two independent surveys. Our proposed method is applied on the two dataset separately and the results obtained are analysed and compared with other approaches reported in the literature on learning pairwise preference models from these datasets.

II. Multivariate Adaptive Regression Spline Model

Multivariate adaptive regression spline (MARS) model is a popular nonparametric model proposed by Friedman to solve regression problems [8]. The key idea is to segment the space of inputs into regions of varying sizes that are fit with linear or cubic splines. Each of the regions has its own regression sub-model and the size of the regions is adjusted by the model according to the nature of the input space; the greater the density of the training data and the greater the complexity of the relationship between the input and output variables, the smaller the regions become. In this sense, MARS combines the strengths of regression trees and spline fitting approaches through the use of piecewise linear basis functions instead of the step functions usually employed in regression trees. Practically speaking, MARS shares the attractive properties of being exceptionally fast to analyse and easy to understand.

MARS attempts to fit adaptive non-linear regression to define relationships between a response variable and some set of predictors through a forward/backward iterative approach [8]. The adaptive non-linear regression uses piecewise basis functions (also known as terms) that are defined in pairs, using a knot or value of a variable that defines an inflection point along the range of a predictor. These knots (parameters) are also determined from the data.

Practically speaking, suppose we have the input data $X = [x_1....x_n]$ where n is the size of the input space, MARS is defined as:

$$y = f(X) = \beta_0 + \sum_{i=1}^{M} \beta_i f_i(X) \tag{1}$$

The model is a weighted (by β_i) sum of basis functions $f_i(X)$ and M is the number of the basis functions. The function $f(X)$ is defined as:

$$f_i(X) = \prod_{k=1}^{K} h(X_{(v,k)}) \tag{2}$$

where X_v is the variable with index v and k is the number of terms in the function ($k \in [1..k_{max}]$).

For order of interactions K=1, the model is additive and for K=2 the model is pairwise interactive.

The MARS algorithm searches over the space of all inputs and predictor values as well as interactions between variables. The method works in two main steps: in the forward stage, an increasingly larger number of basis functions are added to the model. The model selects the knot and its corresponding pair of basis functions that maximise a least squares goodness-of-fit criterion. Knots are chosen automatically and can be placed at any position within the range of each output to define a pair of basis functions. As a result of these operations, MARS automatically determines the most important independent variables as well as the most significant interactions among them. As this might yield an over-fitted model, a backward procedure is then applied; the model is pruned by removing those basis functions that are associated with the smallest increase in the goodness-of-fit. This is done using the *Generalised Cross Validation* (GCV) error which is a measure of the goodness of fit that takes into account both the residual error and the model complexity. This measure is calculated as:

$$GCV = \frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{(1 - \frac{\hat{c}}{n})^2} \tag{3}$$

where n is the number of samples in the training data and $\hat{c}$ is the effective number of parameters and is calculated as:

$$\hat{c} = c + \frac{p * (c-1)}{2} \tag{4}$$

where c is the number of independent basis function, and p is the penalty of adding a basis function.

There are a number of basis functions that are usually defined. The most used are the hinge function and product of a set of hinge functions which are defined as:

$$h(x) = max(0, x - t) | max(0, t - x) \tag{5}$$

where t is a constant called $knot$. The $knot$ value defines the "pieces" of the piecewise linear regression which is also determined from the data.

III. Preference Learning

Preference learning has received increasing attention in the machine learning literature in recent years [1]. The ranking problem has been categorised into three main types, namely: label ranking, instance ranking and object ranking [1]. We focus on object ranking in this paper. Within object ranking, the goal is to learn a ranking function $f(.)$ that produces a ranking of a given subset of objects given their pairwise preferences. More formally, given a set of instances Z and a finite set of pairwise preferences $x_i \succ x_j; (x_i, x_j) \in Z$ x Z, find a ranking function $f(.)$ that returns the ranking of this set Z where $f(x_i) > f(x_j)$. Here, $x_i \succ x_j$ means that instance x_i is preferred to x_j.

Various methods have been presented in the literature for the task of object ranking. Methods based on large-margin classifiers [13], Gaussian processes [14], [4], [7], and neuroevolution [5] have been investigated to learn the ranking function. Neuroevolutionary preference leaning proved to have a powerful approximation capability and to build efficient models of player experience in similar setups to the one at hand [15], [16], [17], [5]. Other supervised learning methods such as standard backpropagation [18], rank support vector machine [19], Cohen method [20], linear preference learning [21] and pairwise preference leaning [1] have also been

employed to learn pairwise preferences with various success. There exists a number of other attempts where the problem of pairwise preference learning is converted into learning a global classifier and therefore standard ranking method can be applied [1]. This paper introduces a new approach for learning pairwise preferences and presents two test cases where the suggested method demonstrated efficient learning and modelling capabilities.

IV. Evolving MARS Models through Grammatical Evolution

Grammatical Evolution (GE) is an evolutionary algorithm based on Grammatical Programming (GP) [22]. The main difference between GE and GP is the genome representation; while a tree-based structure is used in GP, GE relies on a linear genome representation. Similar to general Genetic Algorithms (GAs), GE applies fitness calculations for every individual and it applies genetic operators to produce the next generation.

The population of the evolutionary algorithm is initialised randomly consisting of variable-length integer vectors; the syntax of possible solution is specified through a context-free grammar. GE uses the grammar to guide the construction of the phenotype output. The context-free grammar employed by GE is usually written in Backus Naur Form (BNF). Because of the use of a grammar, GE is capable of generating anything that can be described as a set of rules such as mathematical formulas [23], programming code, game levels [24] and physical and architectural designs [25]. In this paper, we focus on the problem of evolving accurate MARS preference models through the use of GE.

Each chromosome in GE is a vector of codons. Each codon is an integer number used to select a production rule from the BNF grammar in the genotype-to-phenotype mapping. A complete program is generated by selecting production rules from the grammar until all non-terminal rules are mapped. The resultant string is evaluated according to a fitness function to give a score to the genome. In this paper, a Design Grammar (DG) is defined to specify the structure of possible solutions (MARS models in our case) as can be seen in Fig. 1. The DG is defined in a way that allows the construction of model's trees where each expression node represents a possible basis function. A MARS model is constructed by creating a basis function and adding it to the model. According to the grammar, the basis function can be a hinge function or a multiplication of two or more functions. An option of adding an empty node is also added to the grammar to facilitate model simplification by deleting some of the already added functions through the mutation operator.

A. The Fitness Function

The goodness of the models evolved is evaluated with a fitness function. The GCV measure (described in Section II) is usually used for this purpose. However, as we are dealing with pairwise preference data, and since our target values are defined on pairs of instances rather than individuals, we need to revise the definition of GCV. The measure we define, named *Pairwise Generalised Cross Validation* (PGCV), uses the notion of agreement between the model's outputs of a pair of instances and the actual pairwise preference expressed by

```
<Tree> := <Constant> + <Exp>
<Exp> :=  <Constant> * <BasisFunction> + <Exp>
                | <Constant> * <BasisFunction>
<BasisFunction> := <HingeFunction> * <BasisFunction>
                | <HingeFunction>
                | <Empty>
<HingeFunction> := max(0,<Feature> - <Knot>)
                | max(0,<Knot> - <Feature>)
<Feature> := feature1 |  feature2 | ...
<Constant> := [min, max]
<Knot> := [0,1]
<Empty> := []
```

Fig. 1. A simplified version of the grammar employed to specify the structure of MARS model.

the users, i.e. the output given to the preferred instance in a pair should be higher than that given to the unpreferred instance. For this purpose, instead of using the residual sum of squares to calculate the error, we used Kendall tau distance. Practically speaking, the PGCV is calculated as:

$$PGCV = \frac{\sum_{i=1}^{n} u(y_{i.A}, y_{i.B})}{(1 - \frac{\hat{c}}{n})^2} \tag{6}$$

$$u(a,b) = \begin{cases} 1 & \text{if } f(a) <= f(b) \\ 0 & otherwise \end{cases} \tag{7}$$

where n is the number of pairs in the dataset and $f(.)$ is the model output.

V. Experiments

In order to evaluate the proposed approach we use two datasets containing users' subjective feedback collected as pairwise preferences while interacting with a digital system.

A. Dataset 1: Player Data in Super Mario Bros

This dataset contains rich information about game content, player behaviour, and self-reports of hundreds of players playing a modified version of an open source clone of the popular game *Super Mario Bros* (SMB). Subjects were asked to play pairs of game levels and report their affective state using the 4-alternative forced choice (4-AFC) experimental protocol. The data collected allows extraction of several statistical, temporal and spatial features. For the purpose of this study, we use a set of statistical features capturing information about the occurrences and frequencies of selected events. The full set of features considered contains 30 different features that are explained in details in [6].

One of the primary reasons for choosing the Mario dataset is because of its size (780 different pairs) that permits meaningful exploration of the method capabilities. The dataset also offers rich information in terms of the features collected (gameplay and content features). The size of the input space, combined with the complex nature of the dataset (the data is mostly non-linear and the relationship between the features collected and the users' reports can not be easily captured using simple linear or non-linear model) makes the construction of accurate models a challenging task. The dataset also offers the possibility of exploring the strength of the modelling approach in different settings as it contains information about

players' reports of three different emotional states, namely: *engagement, frustration* and *challenge*.

The dataset has been the subject for a number of experiments for modelling player experience [6], [26] and the studies conducted demonstrated the difficulty in learning such models especially if transparency and understandability of the modelling approach are important factors. This makes the dataset and the problem of modelling player experience still interesting as it introduces unsolved research questions.

B. Dataset 2: Expressed Emotions in Music

The second dataset we use consists of pairwise comparisons of 20 different excerpts, each excerpt is of 15 seconds length music piece taken from the middle of selected tracks from the USPOP20022 dataset [27]. The 20 excerpts were chosen such that a linear regression model maps five excerpts into each quadrant of the two-dimensional *arousal-valence* space [28]. Thirteen participants listened to each pair of excerpts and evaluated the emotional dimensions of valence and arousal using the 2-Alternative Forced Choice (2-AFC) experimental paradigm. This dataset has been the subject for a number of experiments for modelling user experience using pairwise preference methods [29], [7].

In order to run machine learning techniques that learn the relationship between the pieces of musics played and the reported affects, representative features are extracted from each excerpt. Several features for modelling the expressed emotions in music have been proposed in the literature [7]. In this work, we used the ones suggested in [29], which are the Mel-Frequency Cepstral Coefficients (MFCCs). The resultant set contains 40 features related to the mean and variance of extracted MFCC parameters (the detailed procedure of extraction is explained in [29]).

The use of this dataset in our experiments facilitate exploration of the generalisable capability of the suggested approaches. The problem of modelling expressed valence and arousal from this dataset is also a challenging task according to the difficulty in learning accurate models [29], [7].

VI. Experimental Setup

The full data in both datasets is used to construct models independently. Since the 4-AFC protocol was used in the first dataset, it was preprocessed to remove the instances with unclear preference (those that elicit non differentiable affective states according to players' reports). The sizes of the resultant datasets are 597, 531 and 629 pairs for reported engagement, frustration and challenge, respectively. The second dataset consists of 190*13 pairs for each emotional dimension (valence and arousal).

There is a fundamental difference in the data collection strategy followed in our testbed cases that influences our decision when building the models. The experimental setup in the second dataset demands evaluation of all music excerpts by each participant. This permits training and testing the models on user-specific data which is not possible in the first dataset since each player provides feedback for only few pairs.

Since we wanted to follow the same modelling strategy to provide a fair comparison of the results, the Leave-One-Out (LOO) strategy is used to evaluate the model generalisability. This can be applied in a straightforward manner in the music dataset because we can simply identify the instances ranked by the same user and use them as one of the fold to leave out. In order to apply this strategy on the Mario dataset, we distinguish between the two main types of the features collected: content (features related to the how the game levels are created) and gameplay (features capturing players' characteristics). This allows us to identify game-dependent instances (those where a set of players played the same game) and consequently apply the LOO paradigm on those.

The modelling accuracy is calculated on the testing sets and the experiment is repeated 20 times. The results reported are the averages over the best generated models from each run. The results obtained are compared with upper and lower baseline models defined as suggested in [7].

The open source GEVA software [30] is used as a core to implement the needed functionalities. The experimental parameters used are the following: 20 runs each ran for 200 generations with a population size of 100 individuals, the ramped half-and-half initialisation method. Tournament selection of size 3, int-flip mutation with probability 0.05, one-point crossover with probability 0.8, and 0 maximum wraps were allowed. All parameters were assigned experimentally.

For model generation, the number of basis functions is bounded between 6 and 30. The penalty value of the $PGCV$ is set to one. For simplification, understandability, and visualisation constrains, we allow only two multiplications of the basis functions ($K = 2$) which facilitates analysis of interactions between two variables.

VII. Analysis

Models are constructed for the two datasets under investigation and in what follows we provide a number of analysis of the models constructed in terms of accuracy, complexity and expressivity and we examine the results.

A. Modelling Accuracy

Using the approach proposed, we were able to construct accurate models for both datasets.

1) Player Data in Super Mario Bros: The results obtained for modelling player experience in SMB for the three emotional states are presented in Table I. The best modelling accuracy is obtained for the prediction of frustration (78.04%) followed by challenge (74.84%) while engagement was the hardest to predict (66.81%). Compared to the baselines, the statistical analysis showed that the proposed model is significantly better than the lower baseline ($p - value < 0.05$). The upper baseline model is slightly better than the constructed model, as expected, with no significant difference. Figure 2 presents the box-plot of the average and standard deviation values of the accuracies of the all models constructed for the three emotional states.

2) Emotion in Music Dataset: High accuracies are also obtained for predicting valence and arousal for the music dataset. The results, presented in Table I, show that average accuracies of 88.56% and 82.92% are obtained for predicting arousal and valence, respectively, which are significantly better

TABLE I. AVERAGE ACCURACIES OBTAINED FROM 20 RUNS OF THE EXPERIMENTS FOR EVOLVING MARS. THE ACCURACIES OF THE BASELINES ARE ALSO PRESENTED FOR COMPARISON.

	Super Mario Bros			Music	
	E	F	C	Arousal	Valence
MARS	66.81%	78.04%	74.84%	88.56%	82.92%
$Baseline_{lower}$	55.14%	62.62%	60.66%	71.84%	69.57%
$Baseline_{upper}$	67.19%	78.75%	75.11%	89.65%	83.85%

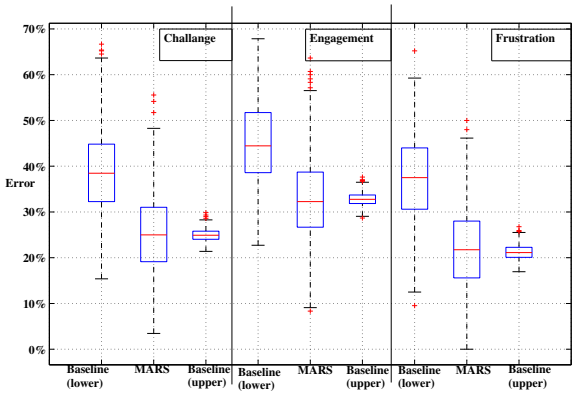

Fig. 2. Average and standard deviation accuracies obtained from 20 runs of the experiments for evolving MARS models for predicting *engagement*, *frustration* and *challenge* from player data in Super Mario Bros dataset. The accuracies of the upper and lower baseline models are also presented for comparison.

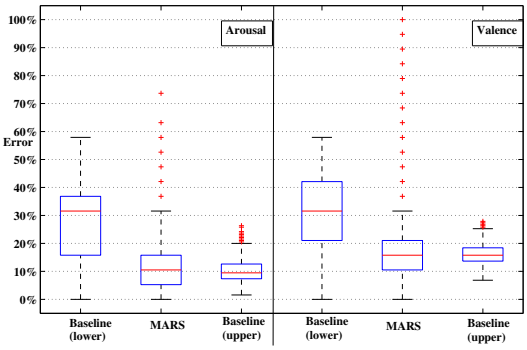

Fig. 3. Average and standard deviation accuracies obtained from 20 runs of the experiments for evolving MARS models for predicting *valence* and *arousal* in the music dataset. The accuracies of the upper and lower baseline models are also presented for comparison.

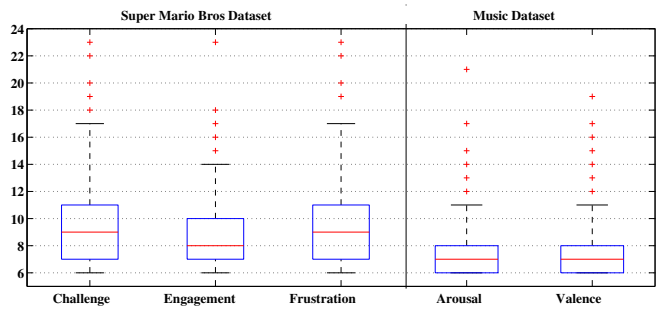

Fig. 4. Average and standard deviation values for the number of features selected by the best models in 20 runs for predicting the expressed emotions.

than the lower baseline ($p-value < 0.05$). Figure 3 represents the values obtained in a box-plot format.

B. Model Complexity

Another interesting aspect to look at is the complexity of the constructed models. This helps us better understand the problem space and the difficulty of modelling it. Several measures could be calculated for this purpose including the convergence speed, analysing the structure of the best models evolved and analysing the parameters of the basis functions. In what follow, we provide a preliminary analysis of complexity by investigating the structure of the models and their basis functions.

1) Number of Selected Features: Figure 4 presents the number of features selected by the best models evolved to predict the reported emotions. As can be seen, most of the models are of average complexity as a relatively low number of features are selected. It is worth noting that a lower number of features is selected to predict valence and arousal in the music dataset compared to the average number of features selected for predicting emotions in the SMB dataset.

2) Relationship Complexity: The complexity of the relationship between the inputs and the predicted emotional states can be analysed by plotting the number of the basis functions used for modelling. Figure 5 presents the results obtained. In general, it seems that the number of basis functions required to construct accurate models is a relatively low. The number, however, seems lower for predicting emotions in the music dataset than that required in the SMB dataset.

3) Feature Importance: In order to provide further analysis of the importance of the features for predicting an emotion, we counted the number of occurrences of each input feature in the final models evolved. Figures 6 and 7 show the percentage of selecting each feature for prediction in the best models evolved for SMB and the music dataset, respectively. The figures indicate a clear significance for some of the features. The number of deaths (*f24*), for instance, seems to have a great impact on how frustrating and challenging the player felt while it has no significant effect on engagement. The same argument can be applied to predicting emotion in the music dataset (Figure 7). One can also notice that the mean values of some of the MFCC features have the heights percentages for both arousal and valence. It is interesting to note that some of the features were very rarely selected indicating their irrelevance for the prediction.

C. Expressivity Analysis

One of the advantages and main motivation behind using MARS as a modelling technique is that it demonstrates powerful regression ability while preserving the expressive power being easy to interpret. In order to further analyse this point and to better understand the complexity of our problem, we investigate the best models constructed using a visualisation method. This is done by plotting the relationship between the variable of the basis functions chosen for evolving the best MARS models and the models' prediction. If a basis function is of $K = 2$, then the relationship is visualised in 3D.

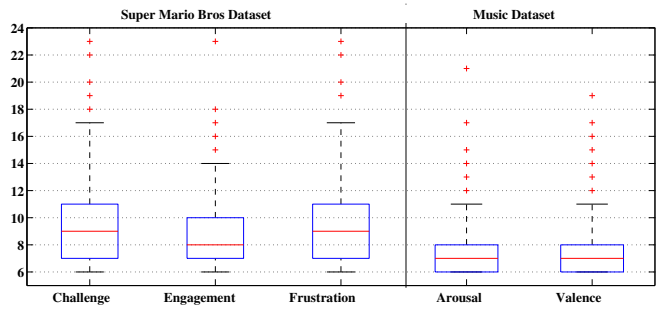

Fig. 5. Average and standard deviation values for the number of functions used for predicting the emotional states in Super Mario Bros and the music dataset.

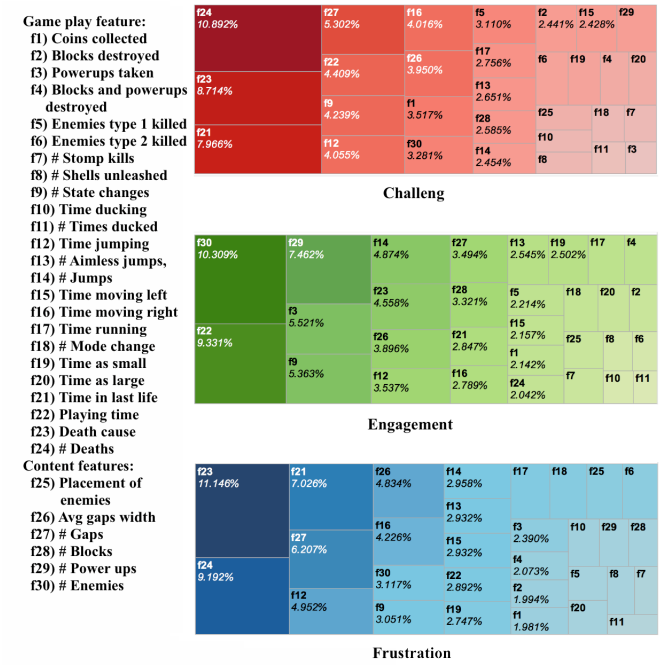

Fig. 6. Percentages of the selection of features in the best models evolved for predicting *challenge, engagement* and *frustration* in SMB dataset. Percentages lower than 1.7 are not shown.

1) Player Data in Super Mario Bros: Figures 8, 9 and 10 show the relationships between each feature selected and the output of the best models for each emotional state in the SMB dataset. As can be seen, the features selected, the complexity of the interaction between them and the complexity of the models varies among the three emotional states. The best model evolved for predicting engagement, for instance, has six basis functions, three of which are multiplications of two hinge functions. The best model for predicting frustration on the other hand has four multiplication functions indicating a more complex interaction between the features. Challenge appears to have the simplest models in term of feature interaction (two multiplication functions).

By looking at the graphs of the individual features selected, we can easily understand the effect of the features on the prediction of the corresponding emotional state. It is clear from the figures that several relationships are of nonlinear nature. Engagement, for example, is a function of the number of coins

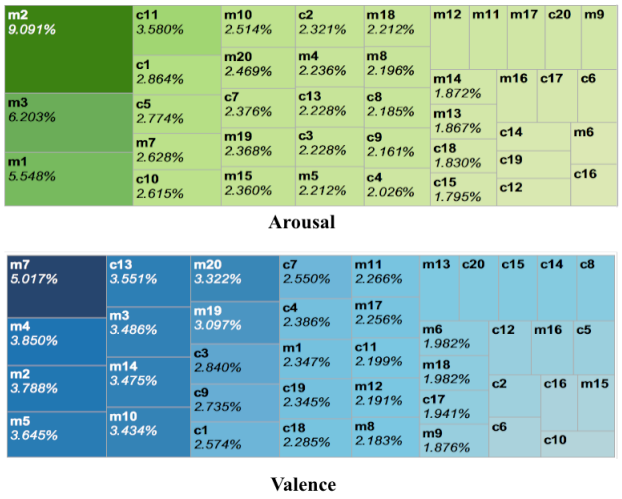

Fig. 7. Percentages of the selection of features in the best models evolved for predicting *valence* and *arousal* in the music dataset. Percentages lower than 1.7 are not shown.

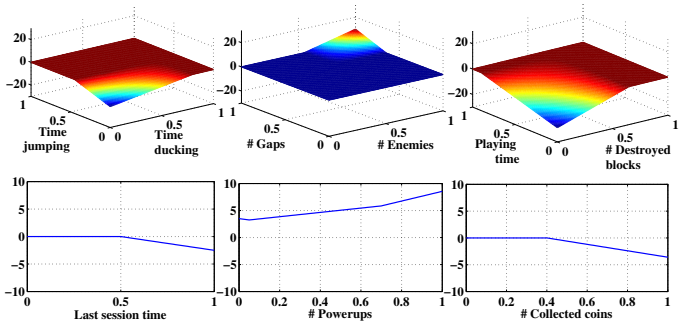

Fig. 8. Features selected by the best MARS model for predicting *engagement* with reference to the prediction accuracy.

collected, the number of power ups presented, the amount of time played in the last session and it is also a function of the interaction between the time spent jumping and ducking, the number of gaps and enemies presented and the number of interactions with the blocks given the amount of time spent playing. The nature of each of these interactions is depicted by the exact characteristics of the hinge functions evolved and the weights assigned to them.

The figures tell, for instance, that engagement is positively correlated with the number of power ups and that it decreases if the session length is more than a certain threshold. The figures also show that adding extra coins above a certain level affects engagement negatively. All of these relationships are captured with one or more hinge functions and they are relatively easy to interpret.

Interactions between the features are captured through multiplications of two (or more, but in our case we limited $k = 2$) hinge functions. Interesting examples of such relationship can be seen in the interaction between the number of gaps and enemies and its effect on engagement. The figure shows that engagement is positively correlated with the interaction between these two factors indicating that this emotional state is affected by the balance between these factors and not only

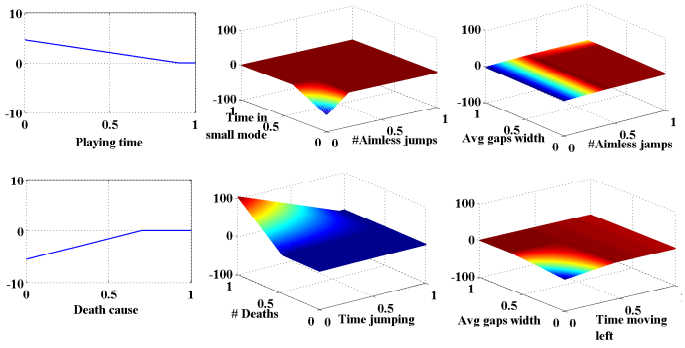

Fig. 9. Features selected by the best MARS model for predicting *frustration* with reference to the prediction accuracy.

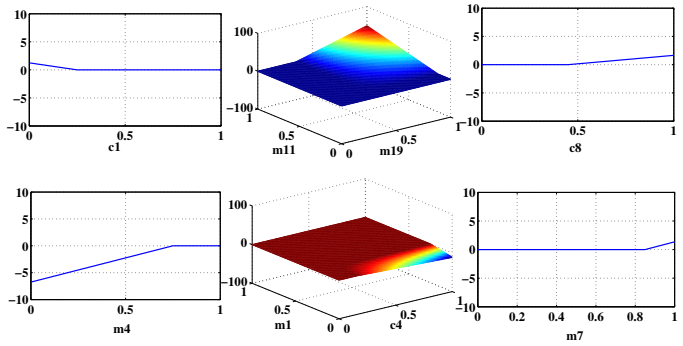

Fig. 11. Features selected by the best MARS model for predicting *arousal* with reference to the prediction accuracy.

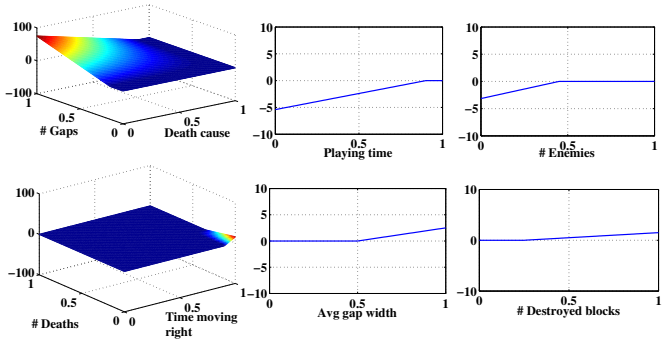

Fig. 10. Features selected by the best MARS model for predicting *challenge* with reference to the prediction accuracy.

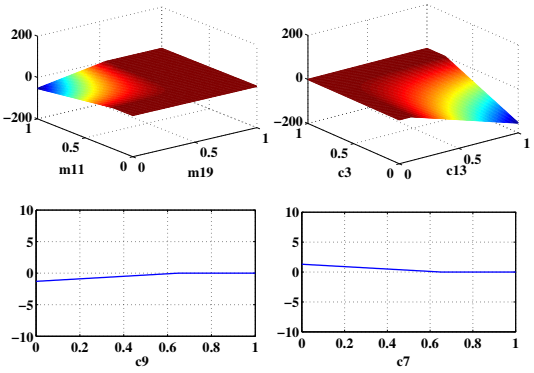

Fig. 12. Features selected by the best MARS model for predicting *valence* with reference to the prediction accuracy.

their absolute values. Another example from the best model of predicting frustration illustrates that a high level of frustration is experienced when the player dies frequently and does not spent much time jumping. This can be explained by a novice player behaviour who is probably not in full knowledge or control of the game mechanics.

2) Emotion in Music Dataset: The relationships between the MFCC features and arousal and valence taken from the best MARS models evolved can be seen in Figures 11 and 12. In split of the fact that assigning a meaningful cognitive interpretation to the acoustic features used for emotion prediction in music is not obvious, what is interesting is to look at the features selected by the models and their dependencies. The results suggest that a small subset of features is important for predicting valence and arousal with high accuracies. Relatively simple models with six and four features out of forty are selected for modelling arousal and valence, respectively. Certain interactions between specific features are important such as the interaction between $m11$ and $m19$ which impacts arousal and the interaction between $c3$ and $c13$ which affects the prediction of valence.

D. Comparison with Previous Attempts

The datasets used are employed in previous studies to construct predictable models of emotions. As explained earlier, the main motivation behind this work is to propose an alternative approach that supports expressiveness while preserving performance. To demonstrate the efficiency of the method proposed, we compared the results obtained with the best ones reported in the literature. The analysis showed superior accuracies for predicting all emotional states in both datasets (compared to [7], [31]). This supports our claim about the efficiency of the suggested approach and motivates further investigation of its capabilities.

VIII. Conclusions and Future Directions

In this paper we proposed a new method for learning pairwise preferences through evolving Multivariate Adaptive Regression Spline Model. We demonstrated the challenges of such task and discussed the attractive characteristics of a good predictor. We presented MARS and illustrated the motivation for choosing it as a modelling approach. We described the traditional method employed to learn a MARS model and we showed that modifications are required if we are to apply this method for preference learning. Because of the nature of the pairwise preference data — being transitive, subjective and lack of a target output per instance— an evolutionary-based method is proposed for training. For this purpose, grammatical evolution is used to define the structure of the models and to evolve them and the definition of the error function is generalised to handle pairwise data. The proposed approach is tested on two relatively large datasets consisting of user and content features and user's feedback of five emotional states (*engagement, frustration, challenge, valence and arousal*) provided as pairwise comparison between two interaction instances. Several experiments were conducted to evaluate the method performance, the modelling complexity and expressiveness. The results showed very promising results

in term of the accuracies obtained and demonstrated powerful expressive characteristics of the models evolved.

The investigations performed in this paper showed that, in general, arousal and valence are easier to predict as illustrated by the accuracies obtained and the complexity of the models evolved (in terms of the number of features selected and the number of the basis functions chosen). Further investigation of this matter is required, preferably on other datasets and domains, to examine whether this is a general trend.

The setup used for all experiments conducted in this paper focuses on optimising the accuracies of predicting one emotional state. In real word application, however, it is usually interesting to predict the appeal of a piece of software along multiple dimensions. It is therefore interesting to investigate the applicability of the proposed method with multi-objective optimisation approaches when, for instance, one could construct models for accurate prediction of two, or more, emotional states (predicting that a level is engaging and challenging at the same time).

IX. Acknowledgement

The research was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This work also was supported in part by the Danish Research Agency, Ministry of Science, Technology and Innovation; project "PlayGALe" (1337-00172). This publication only reflects the authors views. Special thank for Jens Madsen, Bjørn Sand Jensen, and Jan Larsen for providing the emotion music dataset.

References

[1] J. Fürnkranz and E. Hüllermeier, *Preference learning*. Springer-Verlag New York Inc, 2010.

[2] W. W. C. R. E. Schapire and Y. Singer, "Learning to order things," in *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, vol. 10. MIT Press, 1998, p. 451.

[3] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *Machine Learning*. Springer, 2003, pp. 145–156.

[4] W. Chu and Z. Ghahramani, "Preference learning with gaussian processes," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 137–144.

[5] G. N. Yannakakis, M. Maragoudakis, and J. Hallam, "Preference learning for cognitive modeling: a case study on entertainment preferences," *IEEE Transactions on Systems, Man, and Cybernetics. Part A*, vol. 39, pp. 1165–1175, November 2009.

[6] N. Shaker, G. N. Yannakakis, and J. Togelius, "Crowdsourcing the aesthetics of platform games," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 5, no. 3, pp. 276–290, 2013.

[7] J. Madsen, B. S. Jensen, and J. Larsen, "Predictive modeling of expressed emotions in music using pairwise comparisons," in *From Sounds to Music and Emotions*. Springer, 2013, pp. 253–277.

[8] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67, 1991.

[9] C.-J. Lu, T.-S. Lee, and C.-M. Lian, "Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks," *Decision Support Systems*, vol. 54, no. 1, pp. 584–596, 2012.

[10] W. Dwinnell, "Exploring mars: an alternative to neural networks," 1991.

[11] T. Hill and P. Lewicki, *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc., 2006.

[12] D. Rogers, "G/splines: A hybrid of friedman's multivariate adaptive regression splines(mars) algorithm with holland's genetic algorithm," 1991.

[13] C. Fiechter and S. Rogers, "Learning subjective functions with large margins," in *Stanford University*. Morgan Kaufmann Publishers, 2000, pp. 287–294.

[14] M. Gervasio, M. Moffitt, M. Pollack, J. Taylor, and T. Uribe, "Active preference learning for personalized calendar scheduling assistance," in *Proceedings of the 10th international conference on Intelligent user interfaces*, vol. 5. Citeseer, 2005, pp. 90–97.

[15] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 54–67, 2010.

[16] ——, "Modeling player experience in super mario bros," in *CIG'09: Proceedings of the 5th international conference on Computational Intelligence and Games*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 132–139.

[17] H. Martinez, A. Jhala, and G. Yannakakis, "Analyzing the impact of camera viewpoint on player psychophysiology," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[18] G. Tesauro, "Connectionist learning of expert preferences by comparison training," in *Advances in neural information processing systems 1*. Morgan Kaufmann Publishers Inc., 1989, pp. 99–106.

[19] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, vol. 1. IET, 1999, pp. 97–102.

[20] W. W. Cohen, R. E. Schapire, Y. Singer *et al.*, "Learning to order things," *J Artif Intell Res*, vol. 10, pp. 243–270, 1999.

[21] T. Runarsson and S. Lucas, "Imitating play from game trajectories: Temporal difference learning versus preference learning," in *IEEE Conference on Computational Intelligence and Games (CIG)*, 2012, pp. 79 –82.

[22] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 4, pp. 349–358, 2001.

[23] I. Tsoulos and I. Lagaris, "Solving differential equations with genetic programming," *Genetic Programming and Evolvable Machines*, vol. 7, no. 1, pp. 33–54, 2006.

[24] N. Shaker, M. Nicolau, G. Yannakakis, J. Togelius, and M. ONeill, "Evolving levels for super mario bros using grammatical evolution," *IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 304–311, 2012.

[25] M. O'Neill, J. Swafford, J. McDermott, J. Byrne, A. Brabazon, E. Shotton, C. McNally, and M. Hemberg, "Shape grammars and grammatical evolution for evolutionary design," in *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*. ACM, 2009, pp. 1035–1042.

[26] N. Shaker, G. Yannakakis, J. Togelius, M. Nicolau, and M. ONeill, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Transactions on System Man and Cybernetics; Part B: Special Issue on Modern Control for Computer Games*, pp. 1519–1531, 2012.

[27] A. Berenzweig, B. Logan, D. P. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, vol. 28, no. 2, pp. 63–76, 2004.

[28] J. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[29] J. Madsen, B. S. Jensen, J. Larsen, and J. B. Nielsen, "Towards predicting expressed emotion in music from pairwise comparisons," in *9th Sound and Music Computing Conference (SMC) Illusions*, 2012.

[30] M. O'Neill, E. Hemberg, C. Gilligan, E. Bartley, J. McDermott, and A. Brabazon, "Geva: grammatical evolution in java," *ACM SIGEVOlution*, vol. 3, no. 2, pp. 17–22, 2008.

[31] N. Shaker, G. N. Yannakakis, and J. Togelius, "Towards player-driven procedural content generation," in *Proceedings of the 9th conference on Computing Frontiers*. ACM, 2012, pp. 237–240.